Customer Profile Insights

描述性分析

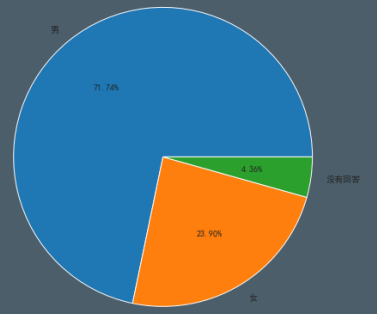
```
#### 数据缺失值百分比 ####
车主id          0.000000
姓名            0.000000
性别            0.000000
生日           47.599223
教育程度        41.659728
职位           37.246739
行业           45.628643
家庭年收入      59.089648
家庭成员人数    72.356370
是否拥有驾照    74.687760
手机            0.000000
省份            0.000000
城市            0.000000
dtype: float64
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 3955 entries, 1-8MN-68814 to 1-3S29-22819
Data columns (total 16 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   车主id      3955 non-null   object
 1   姓名        3955 non-null   object
 2   性别        3947 non-null   object
 3   生日        1965 non-null   datetime64[ns]
 4   教育程度    2306 non-null   object
 5   职位        2473 non-null   object
 6   行业        2152 non-null   object
 7   家庭年收入  1588 non-null   object
 8   家庭成员人数  1138 non-null   object
 9   是否拥有驾照  1059 non-null   object
 10  是否大客户  30 non-null     object
 11  手机        3954 non-null   float64
 12  省份        3693 non-null   object
 13  城市        3611 non-null   object
 14  址区        3086 non-null   object
 15  邮编        2647 non-null   object
```

客户基本数据有3955条，16个特征维度，其中7个特征严重缺失，经过缺失值填补和特征转换，最终的到3602条有效数据，以及筛选出10条相关性较高的特征。

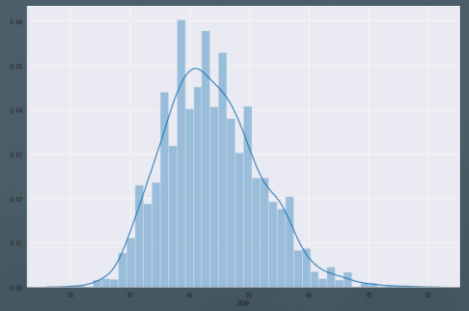
车主ID	性别	教育程度	职位	行业	家庭年收入	家庭成员人数	是否拥有驾照	省份	城市	年龄
1-8MN-68814	男	初中	总裁/总经理/总监/企业高管	家居、装饰	500000.0	4.0	Y	北京市	北京市	43.695031
1-3YPW-6314	男	本科	公司拥有者（老板）/合伙人	家居、装饰	500000.0	4.0	Y	江西省	景德镇市	40.000000
1-8LJ-21917	女	高中	总裁/总经理/总监/企业高管	电气、电器、仪器制造行业	100000.0	3.0	Y	河北省	衡水市	51.333333
1-8MK-70795	女	高中	总裁/总经理/总监/企业高管	电气、电器、仪器制造行业	230000.0	3.0	Y	安徽省	合肥市	57.000000
1-3S29-22819	男	高中	总裁/总经理/总监/企业高管	电气、电器、仪器制造行业	230000.0	3.0	Y	山东省	菏泽市	37.000000

特征分析



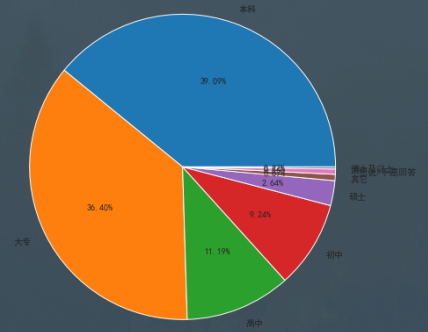
性别

客户主要集中在男性群体，占比达到75.7%；说明我们更应注重男性偏好的相关车型的业务发展。



年龄

客户年龄分布集中在35到50岁，峰值在40岁左右；说明客户集中在中年人群，更适合推荐中年人偏好的车型。

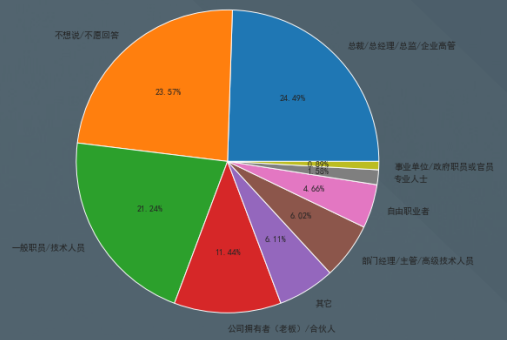


教育程度

38.11%的客户是本科学历，
36.68%的客户是大专学历，
11.42%的客户是高中学历。硕士、博士级以上只占了3%左右；
说明客户的教育程度在中等水平。

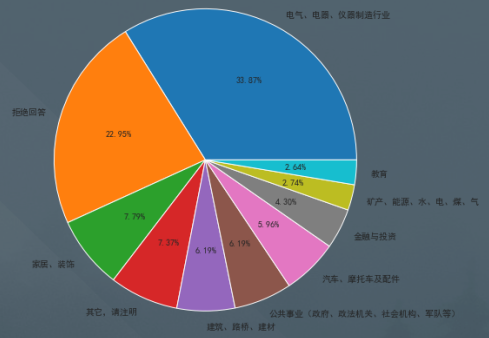
Customer Profile Insights

特征分析



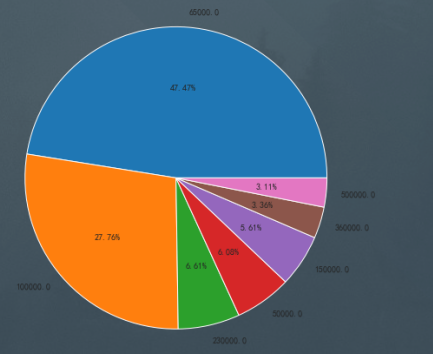
职位

24.5%左右的客户是企业高管等高层职位，有21.23%的客户是一般职员或者技术人员，而事业单位或者正负官员占比是最少的，不到1%；并且客户的职位呈现两级分化，高层职位和低层职位占比相近。



行业

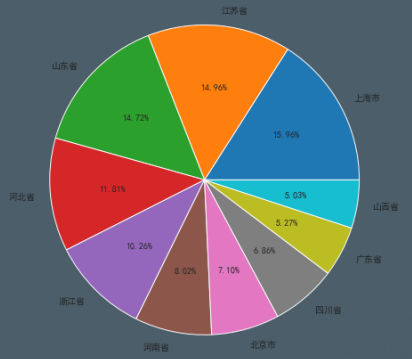
33.85%的客户都集中在电气、电源、仪器制造行业。而其他行业。例如：家居、建筑、汽车、政府、金融、教育等，比重都差距不大，其中也有23%的人拒绝回答自己从事的行业。



教育程度

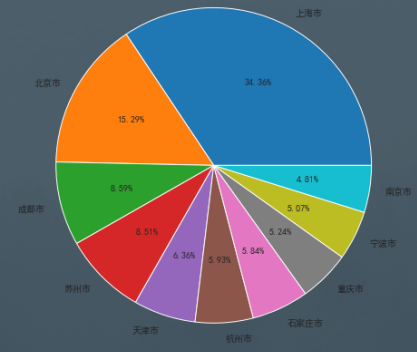
47.5%的客户收入水平在65000元左右，27.7%在8-12万元；年收入5万元以下和年收入在12-18万或18-27万的客户基本持平，都在6%左右；客户普遍集中在中等偏低收入水平的客户，总共占比大概在86%左右（针对5-15万年收入群体）。

特征分析



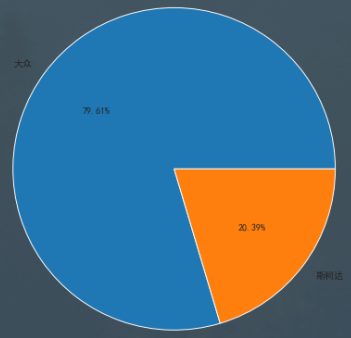
省份

客户在上海市（直辖市）的总数量是最高的，达到15.96%，但是也和江苏省的14.96%、山东省的14.72%相差不多，山西省占比5%是最少的，在南北地区分布来看，客户分布数量基本持平。



城市

大部分的客户都分布在上海市和北京市，总占比达到50%左右。其中上海市占比最高，达到34.36%，北京市排在第二，达到了15.29%。



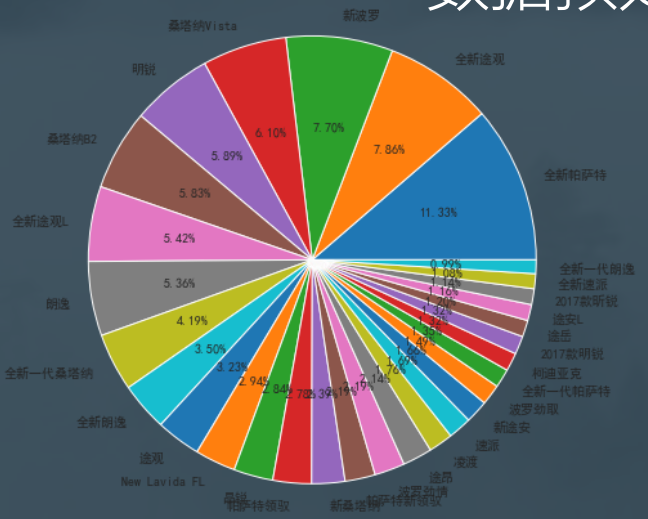
品牌

79.6%的品牌来自于大众；根据工单一级到三级的业务占比，主要都是集中在大众品牌上的俱乐部事宜、维修、保修、配件以及订单问题，而且工单流转分配效率较高。

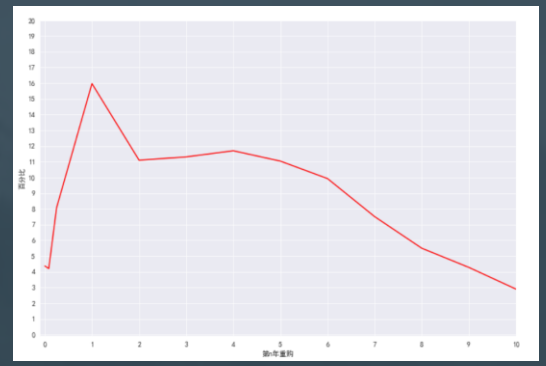
Repurchase Related Insights

数据预处理

#### 数据缺失值百分比 ####	
车主ID	0.000000
VVIN	2.231136
市场车型	0.501379
车型	0.501379
购买日期	0.250689
开票价格	0.000000
ASSET_REF_EXPR	0.501379
OU_NAME	6.442717
OU_ABREV	6.442717
OU_CITY	6.492855
OU_COUNTY	9.275508



	价格差异
count	3.343000e+03
mean	5.846422e+04
std	1.146036e+05
min	-4.815000e+05
25%	0.000000e+00
50%	4.980000e+04
75%	1.265000e+05
max	2.222900e+06



数据概况&缺失值处理

客户重购数据含有7978行和11列，缺失值占比均低于10%，并且有486位客户暂时还未重购，而3343位客户有重购记录。

一共有76中不同款车型，平均价格在183000元上下。

市场车型分析

占据主导地位的车型是全新帕萨特（11.33%）；最贵的车款是途锐，均价在639000元左右，我根据对价格的分析结果对其进行评级（共7个等级），所以针对不同类型的客户，我们根据其基本信息、购买力、消费意愿以及积极程度来预测他们的回购情况，并相应的推荐车款。

重购价差分析

总体来说，重购车辆的价格会比首购平均高出46000元左右，变化量的中位数在7000元左右，最大价格差在220左右，最小则为-48万左右，差异变动很大，但整体成正态分布。

重购时间差分析

上图为重购情况和重购时间差的曲线图；当天重购和当月重购的比率接近（4%），而大部分客户选择在第一年内重购（16%），在第二、三、四、五年里重购的比重也十分相似（11%），之后重购率稳步下降；所以客户集中在短中期（5年以内）回购。

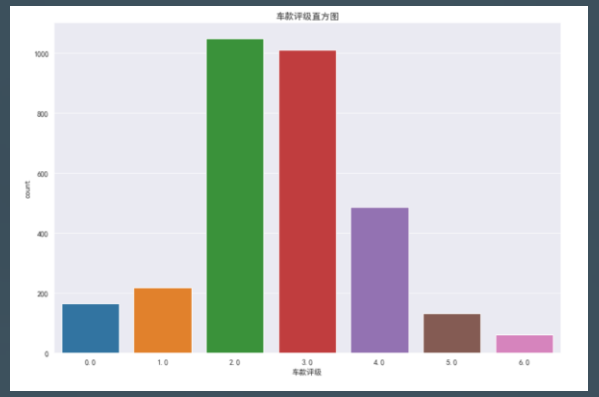
Repurchase Related Insights

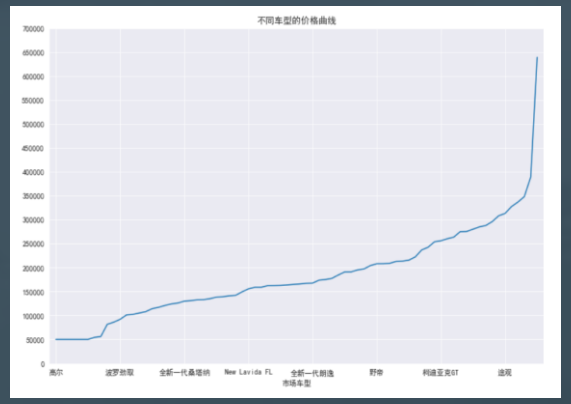
数据分析&挖掘

车主ID	开票价格	价格差异	时间差异	车款评级
1-10LD-2240	256600.0	-8400.0	6.0	4.0
1-10ZA-842	131900.0	-163800.0	510.0	2.0
1-11GN3TJ	119650.0	-1700.0	495.0	2.0
1-11GNTKY	191850.0	-143900.0	3.0	3.0
1-11HMNQQ	323850.0	63900.0	493.0	5.0
...	...	...	...	...
1-Z7U-601	187950.0	24100.0	947.0	3.0
1-ZG3W1L	224900.0	76000.0	716.0	3.0
1-ZPOITB	114900.0	4000.0	39.0	2.0
1-ZUL-1836	192250.0	95100.0	1403.0	3.0
1-ZUZ-86	197880.0	26560.0	172.0	3.0

回购客户特征分析

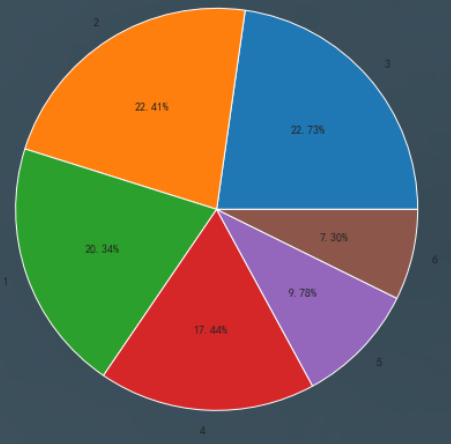
开票价格、回购价格差异、回购时间差异对于车款评级具有决定性因素，车款评级分布如下：





客户活跃度分析

根据客户的消费能力和重购时差来界定客户的活跃程度；65%的客户活跃度较高，下图为分级情况。

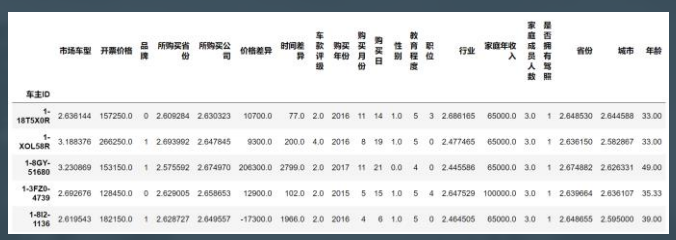


市场车型	3120 non-null object
开票价格	3120 non-null float64
品牌	3120 non-null object
所购买省份	3120 non-null object
所购买公司	3120 non-null object
价格差异	3120 non-null float64
时间差异	3120 non-null float64
车款评级	3120 non-null float64
购买年份	3120 non-null int64
购买月份	3120 non-null int64
购买日	3120 non-null int64
性别	3120 non-null object
教育程度	3120 non-null object
职位	3120 non-null object
行业	3120 non-null object
家庭年收入	3120 non-null float64
家庭成员人数	3120 non-null float64
是否拥有驾照	3120 non-null object
省份	3120 non-null object
城市	3120 non-null object
年龄	3120 non-null float64

客户信息交互

通过连接客户基本信息数据和客户回购信息数据进行信息交互，整合出了3120条有效数据以及21个有效特征；数据概况入下：

车主ID	市场车型	开票价格	品牌	所购买省份	所购买公司	价格差异	时间差异	车款评级	购买年份	购买月份	购买日	性别	教育程度	职位	行业	家庭年收入	家庭成员人数	是否拥有驾照	省份	城市	年龄
1-10LD-2240	全新帕萨特	256600.0	大众	河北省	沧州兴华汽车销售服务有限公司	-8400.0	6.0	4.0	2013	12	16	男	高中	公司拥有者（老板）/合伙人	汽车、摩托车及配件	100000.0	3.0	Y	河北省	沧州市	38.00
1-10ZA-842	全新朗逸	131900.0	大众	江苏省	南京公田发展股份有限公司	-163800.0	510.0	2.0	2015	5	6	女	大专	公司拥有者（老板）/合伙人	电气、电器、仪器制造行业	230000.0	3.0	Y	江苏省	南京市	55.00
1-11GN3TJ	新桑塔纳	119650.0	大众	安徽省	宿州海鹰汽车销售服务有限责任公司	-1700.0	495.0	2.0	2017	7	8	男	高中	公司拥有者（老板）/合伙人	其它，请注明	500000.0	2.0	Y	安徽省	宿州市	60.33
1-11GNTKY	全新一代桑塔纳	191850.0	大众	北京市	北京市艾澜汽车有限公司	-143900.0	3.0	3.0	2016	3	3	女	本科	一般职员/技术人员	电气、电器、仪器制造行业	65000.0	3.0	Y	河北省	沧州市	41.50
1-11HMNQQ	全新途观L	323850.0	大众	贵州省	贵州华通众源汽车贸易服务有限公司	63900.0	493.0	5.0	2017	7	7	男	初中	公司拥有者（老板）/合伙人	建筑、路桥、建材	65000.0	4.0	Y	贵州省	贵阳市	55.67



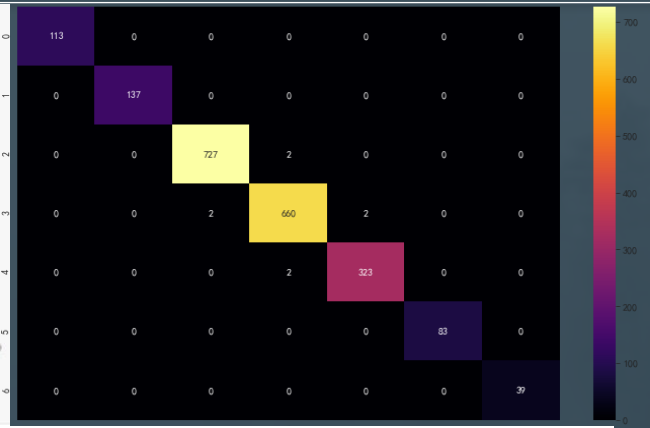
特征工程&相关性分析

对categorical变量进行encoding处理，对其进行适当的编码转换，从而为带入模型做准备；我采取了target encoding以及平滑处理解决了数据泄露以及维度爆炸的问题，并使用label encoding对序列数据进行编码，从而使得所有数据序列化、数字化。

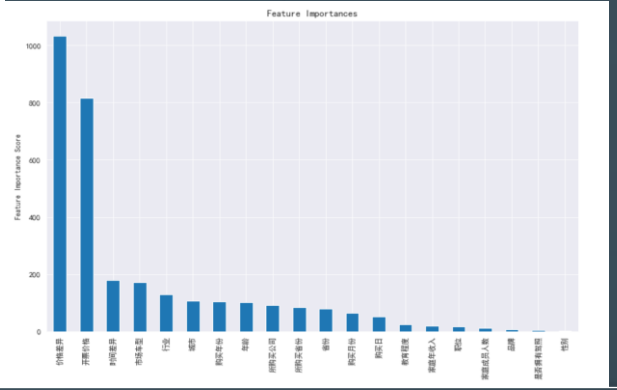
下图为特征之间的关于heatmap的相关性分析图表。

基于XGBoost算法对客户重购的车款评级预测以及车型推荐

```
def modelfit(alg, dtrain, predictors, useTrainCV=True, cv_folds=5, early_stopping_rounds=50, target = '车款评级'):
    '''
    alg: 算法
    dtrain: 训练集包含target
    predictors: 特征column
    '''
    if useTrainCV:
        xgb_param = alg.get_xgb_params()
        xgtrain = xgb.DMatrix(dtrain[predictors].values, label=dtrain[target].values)
        cvresult = xgb.cv(xgb_param, xgtrain, num_boost_round=alg.get_params()['n_estimators'], nfold=cv_folds,
            metrics='mlogloss', early_stopping_rounds=early_stopping_rounds)
        alg.set_params(n_estimators=cvresult.shape[0])

    #Fit the algorithm on the data
    alg.fit(dtrain[predictors], dtrain[target],eval_metric='mlogloss')

    #Predict training set:
    dtrain_predictions = alg.predict(dtrain[predictors])
    dtrain_predprob = alg.predict_proba(dtrain[predictors])[:,1]

    #Print model report:
    print ("\nModel Report")
    print ("Accuracy : %.4g" % metrics.accuracy_score(dtrain[target].values, dtrain_predictions))
    print(confusion_matrix(dtrain_predictions,dtrain[target].values))
    print(classification_report(dtrain_predictions,dtrain[target].values))

    pd.DataFrame(alg.predict_proba(dtrain[predictors])).plot()
    plt.show()

    sns.heatmap(pd.DataFrame(confusion_matrix(dtrain_predictions,dtrain[target].values)), annot=True, cmap='inferno', fmt="d")
    plt.show()


    feat_imp = pd.Series(alg.get_booster().get_fscore()).sort_values(ascending=False)
    feat_imp.plot(kind='bar', title='Feature Importances')
    plt.ylabel('Feature Importance Score')
```



	precision	recall	f1-score	support
0.0	1.00	1.00	1.00	113
1.0	1.00	1.00	1.00	137
2.0	1.00	1.00	1.00	729
3.0	0.99	0.99	0.99	664
4.0	0.99	0.99	0.99	325
5.0	1.00	1.00	1.00	83
6.0	1.00	1.00	1.00	39
accuracy			1.00	2090
macro avg	1.00	1.00	1.00	2090
weighted avg	1.00	1.00	1.00	2090



XGBoost算法介绍

XGBoost是boosting算法的其中一种。Boosting算法的思想是将许多弱分类器集成在一起形成一个强分类器。XGBoost是一种提升树模型，它是将许多树模型集成在一起，形成一个很强的分类器，所用到的树模型则是CART回归树模型。

数据建模

对新数据进行数据分析、特征工程、数据转换以及利用XGBoost算法对数据根据车款评级分类进行建模，并可以通过评级内车款的销量加权得到相应的回购概率，筛选排名前10的车款进行推荐；我使用Python的XGBClassifier来建模（左上图为个人撰写的代码），并采用5折交叉验证来提升模型泛化能力和可信度，通过使其多分类对数损失函数最小化来实现多分类任务。

模型评估

通过对超参数（学习率、树总量、最大深度等）的调节，刻画confusion matrix（右上）和classification report（左下）来体现模型在测试机上的泛化能力，准确度、特异度以及召回率都达到95%以上，说明模型有很强的泛化能力，能对新数据进行预测。并且通过随机森林的特征分裂算法刻画了各特征的重要性（右下）；影响因素前五的特征分别为重购价差、开票价格、重购时差、市场车型以及客户行业。

代码链接：https://zg104.github.io/xgb